УДК 004.8

Ю.Г. Кривонос, Ю.В. Крак, О.С. Загваздін, Г.М. Єфімов
Інститут кібернетики ім. В.М. Глушкова НАН України, м. Київ, Україна
{yuri.krak, alex.zagvazdin}a @gmail.com

Сегментація мовних голосових сигналів за ознакою зміни диктора

Запропоновано підхід до сегментації голосових мовних сигналів за ознакою зміни диктора та способи визначення позицій зміни диктора в голосовому мовному сигналі. Позиції зміни диктора визначаються за допомогою аналізу множин характеристичних векторів в околі паузи на основі Байєсівського інформаційного критерію. Покращення якості характеристичних векторів досягається за допомогою використання сегментів з рівнем енергії не нижче певного порогу. Також пропонується адаптивний підхід для автоматичного визначення пауз у мовному сигналі.

Вступ

Задача сегментації голосових сигналів за ознакою зміни диктора є важливою задачею цифрової обробки мовних голосових сигналів, що використовуються в інформаційних системах зберігання і пошуку мовної голосової інформації, системах автоматизованого комп'ютерного документування [1], системах розпізнавання мовної голосової інформації тощо. В системах зберігання і пошуку мовної голосової інформації сегментація за ознакою зміни диктора дозволяє пов'язувати сегменти звукових сигналів з певними дикторами, що дозволяє відтворювати пошук звукових фрагментів, що пов'язані з певною особою. В системах автоматизованого стенографування така сегментація дозволяє підвищити інтелектуальність розбиття вхідного сигналу на сегменти і створює можливість асоціації сегментів звукової інформації з дикторами під час обробки. В системах розпізнавання мовних сигналів визначення позицій зміни диктора дає можливість налаштовувати систему під акустичні особливості мови певного диктора, дає змогу підвищити якість розпізнавання.

Задача сегментації голосового сигналу за ознакою зміни диктора здебільшого полягає у пошуку позицій у вхідному сигналі, в яких відбувається зміна диктора, за умови, що інформація про кількість дикторів у сигналі чи акустичні характеристики голосів дикторів заздалегідь не відома. Існує досить велика кількість підходів до задачі визначення зміни диктора. Здебільшого такі підходи базуються на порівнянні множин характеристичних ознак в сусідніх ділянках вхідного сигналу. При цьому як характеристичні ознаки використовуються, як правило, вектори коефіцієнтів мел-кепстр. Інколи як додаткові ознаки також використовують енергію сигналу, максимальні значення перетворення Фур'є на ділянці сигналу [2], частоту основного тону (пітч), коефіцієнти лінійного передбачення тощо. Підходи різняться в тому, як визначаються сусідні вікна сигналу, для яких буде проводитись порівняння, та мірами, за якими порівнюються множини характеристичних векторів, що відповідають ділянкам сигналу, що обробляється.

Так, в роботах [3], [4] як міра, за якою порівнюються множини характеристичних векторів, використовується міра дивергенції, що дає змогу реалізації підходу до визначення позиції зміни диктора у реальному часі. Проте, оскільки на кожний момент часу порівнюються лише кілька сусідніх сегментів, побудувати вдалу модель диктора за умови обмеженої інформації важко. В роботі [5] пропонується як міру для порівняння мно-

жин використовувати зважену Евклідову відстань між векторами. За такого підходу випадкові збурення у вхідному сигналі можуть призвести до суттєвого погіршення якості визначення позицій зміни диктора. Для усунення цього недоліку в [6] запропоновано порівнювати характеристичні множини характеристичних векторів на основі попарного порівнювання векторів з різних множин і використовування медіани відстаней між окремими векторами як міри відмінності між множинами. Такий підхід дозволяє зменшити вплив випадкових збурень на якість роботи алгоритму, проте, як показали експерименти, в результаті роботи підходу визначається досить велика кількість помилково визначених позицій зміни диктора (там, де система вважає, що відбувається зміна диктора, а насправді її немає). В роботі [7] пропонується вирішувати задачу про наявність зміни диктора між двома сегментами сигналу як задачу перевірки гіпотези про наявність зміни диктора, за умови, що множини характеристичних векторів є множинами нормально розподілених незалежних векторних випадкових величин. Як критерій до прийняття чи відхилення гіпотези використовується Байєсівський інформаційний критерій. Такий підхід дозволяє досить точно визначити позиції зміни диктора, проте оскільки на кожному кроці роботи алгоритму порівнюються два сусідні сегменти і не враховується наявність пауз у сигналі, а також тому, що як критерій для прийняття рішення про наявність зміни диктора використовуються локальні максимуми Байєсівського інформаційного критерію, підхід може генерувати досить велику кількість помилково визначених змін дикторів.

У даній роботі пропонується стратегія сегментації мовного голосового сигналу за ознакою зміни диктора. Підхід базується на визначенні позиції зміни диктора з використанням логарифмічного відношення правдоподібності і Байєсівського інформаційного критерію, проте на відміну від методів, запропонованих в [7], пропонується шукати позиції зміни диктора лише в околі пауз у голосовому сигналі, що за умови точного визначення пауз дозволяє зменшити ризик помилкового визначення зміни диктора там, де їх немає. Для визначення пауз пропонується підхід, що базується на логарифмічній енергії сигналу з використанням автоматичного адаптивного визначення порогу. У характеристичних векторах пропонується використовувати, окрім коефіцієнтів мел-кепстр, пітч, що дозволяє підвищити точність визначення позиції зміни диктора, коли диктори відрізняються за статтю чи віком.

Постановка задачі

Для сегментації звукового сигналу за ознакою зміни диктора необхідно визначити позиції зміни диктора в сигналі. Покладемо, що зміна диктора відбувається в околі ділянки сигналу, де є пауза: перед тим як закінчує говорити один диктор і починає інший, є певний період мовчання. У справжніх сигналах це не завжди так, оскільки диктори можуть перебивати один одного, проте в такому випадку складно сегментувати сигнал, оскільки подібні ділянки неможливо однозначно віднести до жодного з сегментів. Тому для даної задачі обмеження про те, що зміна диктора відбувається в околі паузи, є припустимим.

Отже, першою задачею є визначення ділянок сигналу, що відповідають паузам. Після того, як такі ділянки знайдені, необхідно певним чином порівняти характеристики звукового сигналу до і після паузи. Нехай $X = \{x_1, x_2, ..., x_n\}$ – множина характеристичних векторів, що відповідають певній ділянці сигналу до паузи, і $Y = \{y_1, y_2, ..., y_n\}$ – множина характеристичних векторів, що відповідає ділянці сигналу після паузи. N_x – кількість векторів у першій множині, N_y – кількість векторів у іншій множині. Характеристичні вектори, які використовуються в задачі, обраховуються на досить короткому вікні сигналу,

сусідні вікна певною мірою перетинаються між собою. Докладніше характеристичні вектори будуть описані нижче. Нехай $Z = X \cup Y$ – об'єднання множин характеристичних векторів, з кількістю точок $N = N_x + N_y$. Множини X і Y порівнюються між собою за допомогою певної міри відмінності, і якщо відмінність між цими множинами є достатньо великою, то робиться висновок про те, що в околі паузи, що знаходиться між цими множинами, відбувається зміна диктора. В такому випадку, при сегментації, сегмент, що відповідає першому диктору, буде закінчуватися на початку даної паузи, а сегмент, що відповідає іншому диктору, буде починатися наприкінці даної паузи.

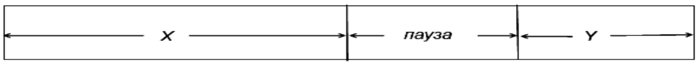


Рисунок 1 – Ділянки сигналу до і після паузи, що порівнюються

Необхідно знайти таку міру відмінності і таке значення порогу для міри відмінності, що при порівнюванні множин перевищення порогу буде відповідати зміні диктора і дозволить максимізувати кількість коректно визначених змін диктора і мінімізувати кількість помилок першого і другого роду (помилково визначених змін диктора і пропущених змін диктора).

Визначення пауз у голосовому сигналі

Оскільки позиції зміни диктора шукаються в околі пауз, необхідно визначити ділянки сигналу, що відповідають паузам. Ця задача сама по собі є нетривіальною, особливо для сигналів з високим чи нестаціонарним рівнем сторонніх шумів. Більшість підходів до визначення пауз базуються на вимірюванні рівня енергії сигналу в ділянці, що обробляється, і порівнянні отриманого рівня з певним чином заданим пороговим значенням. Відрізняються різні підходи між собою тим, яким чином вимірюється енергія сигналу і яким чином задається порогове значення. Як міра енергії використовуються логарифмічна енергія сигналу [8], кількість перетинів нуля (Zero Crossing Rate) [9], дисперсія вимірів сигналу [10] чи комбінації цих метрик.

Для адаптації до нестаціонарного рівня шуму були запропоновані адаптивні підходи до визначення порогових значень [10]. Запропонований нижче підхід є розвитком адаптивного методу визначення пауз, запропонованого в [10], в якому як міра енергії використовується логарифмічна енергія ділянки сигналу:

$$E = 10\log_{10}\left(\sum_{i=1}^{M} x_i^2\right) \tag{1}$$

де x_i – значення сигналу, $i = \{1,2,...,M\}$, M – довжина вікна, що аналізується.

Виміри енергії для послідовно розташованих вікон згладжуються методом медіанного згладження 5 порядку [11]. Це дозволяє уникнути впливу випадкових збурень у сигналі, що, як правило, пов'язані зі сторонніми шумами.

Для визначення порогового значення використовується інформація про попередні 10 с звучання сигналу (покладаємо, що на ділянці в 10 с буде принаймні одна пауза). Для визначення порогу використовується інформація про максимальне і мінімальне значення енергії сигналу на ділянці в 10 с, і рішення про те, що певне вікно відповідає паузі, робиться за наступним критерієм:

$$E < E_{\min} \vee \frac{E - E_{\min}}{E_{\max} - E_{\min}} < 0.2, \tag{2}$$

тут E – рівень енергії сигналу у поточному вікні, $E_{\min}$ – мінімальний рівень енергії на ділянці, $E_{\max}$ – максимальний рівень енергії на ділянці в 10 с.

Для пошуку пауз використовуються вікна тривалістю 20 мс і перетинаються між собою на 10 мс. Вікна з паузами, що розташовані підряд одне за одним, об'єднуються в одну паузу. При визначенні пауз додаткові обмеження також накладаються на мінімальну тривалість паузи, що дозволяє виключити занадто короткі вікна з невеликим рівнем енергії.

Визначення позиції зміни диктора

Задача визначення, чи є в околі певної паузи зміна диктора, формулюється як задача перевірки гіпотези, що порівнює дві гіпотези: H_0 – про те, що зміна диктора в околі даної паузи відсутня, і H_1 – про те, що в околі даної паузи відбувається зміна диктора. Покладемо, що величини, з яких складаються множини X і Y, є незалежними і однаково розподіленими. Тоді параметри нормального розподілу Θ_Z для множини Z, яка є об'єднанням множин X і Y, можуть бути оцінені за допомогою методу максимальної правдоподібності. Логарифмічне відношення правдоподібності для множини спостережень Z за гіпотези H_0 задається таким співвідношенням:

$$L_0 = \sum_{i=1}^{N_z} \log p(x_i \mid \Theta_Z) + \sum_{i=1}^{N_y} \log p(y_i \mid \Theta_Z), \qquad (3)$$

де $p(x \mid \Theta)$ – ймовірність того, що x справджується за умови Θ.

Для перевірки гіпотези H_1 обраховуються параметри нормальних розподілів індивідуальні для наборів спостережень X і Y, які відповідно позначаються як Θ_X і Θ_T. При цьому логарифмічне відношення правдоподібності запишеться як:

$$L_1 = \sum_{i=1}^{N_x} \log p(x_i \mid \Theta_X) + \sum_{i=1}^{N_y} \log p(y_i \mid \Theta_Y). \qquad (4)$$

Звичайна міра відмінності між множинами в такому випадку задається як

$$d_0 = L_1 - L_0. \qquad (5)$$

Оскільки параметри правдоподібності для гіпотези H_1 визначаються окремо для множин X і Y, звідси складові частини в правій частині (4) завжди більше відповідних складових частин (5), то $L_1 \geq L_0$ і $d_0 \geq 0$. Більш надійною мірою відмінності між множинами, проте, є міра, що базується на Байєсівському інформаційному критерії, де значення міри коректується відповідно до кількості елементів в кожній з множин, що порівнюються. Міра відмінності, що базується на Байєсівському інформаційному критерії, обраховується за формулою:

$$d_1 = L_1 - L_0 - \frac{\lambda}{2} \Delta K \log N, \qquad (6)$$

де $\Delta K = N_x - N_y$, λ – це параметр, який теоретично має бути рівним 1,0. Причому такий критерій теоретично дає змогу уникнути визначення порогу для міри відмінності. За такого підходу позиції зміни диктора будуть визначатися як точки, в яких функція різниці між множинами набуває локального максимуму. Проте на практиці відсутність порогового значення не завжди дає оптимальний результат. Як показали експерименти, кількість помилково визначених змін диктора навіть на досить якісному сигналі є достатньо високою при різних значеннях λ. Тому вважаємо, що для прийняття рішення про наявність зміни диктора в околі певної паузи необхідно, щоби значення міри відмінності між множинами характеристичних векторів перевищувало певний поріг, який підбирається вручну для конкретного сигналу.

Як характеристичні вектори, з яких складаються множини X і Y, обрано вектори, що складаються з 13 коефіцієнтів мел-кепстр і пітчу як 14 виміру. Пітч є характеристичною ознакою, що досить точно передає статеві та вікові відмінності між голосами дикторів. При цьому характеристичні ознаки обраховуються по ділянках сигналу тривалістю 30 мс, що перетинаються між собою на 10 мс, до яких застосовані віконні функції Хеннінга. Слід зауважити, що для визначення характеристичних ознак, що відповідають дикторам, доцільно використовувати аналітичні вікна більшої тривалості, ніж зазвичай використовуються для розпізнавання мовних голосових сигналів, оскільки характеристики, що відрізняють дикторів, є більш «розтягнутими» в часі. До того ж такий підхід дозволяє дещо скоротити кількість необхідних обчислень.

Експериментально було також встановлено, що найбільшу кількість інформації про диктора несуть ділянки сигналу з більшою логарифмічною енергією. Тому при знаходженні позицій зміни диктора до розрахунку беруться лише ділянки, в яких логарифмічна енергія перевищує певний поріг. Для експериментів, що проводилися в рамках даної роботи, використовувався поріг в 40 db. Це дозволило не лише досить суттєво скоротити кількість необхідних обчислень, але й підвищити точність визначення позицій зміни диктора за рахунок того, що ділянки з малим рівнем енергії часто несуть «зайву» інформацію, яка не характеризує диктора (як правило, це сторонні шуми).

Границі ділянок сигналу, що порівнюється, визначаються таким чином:

1. Множина X складається з характеристичних векторів, обчислених для ділянок сигналу між попередньою зміною диктора і початком поточної паузи. Якщо кількість елементів в множині X становить більше 200, вектори, що відповідають більш раннім ділянкам сигналу, виключаються з розрахунків, для того щоб тримати кількість обчислень, що необхідно робити на кожному кроці алгоритму, в раціональних межах.

2. Множина Y складається з характеристичних векторів, обрахованих на ділянці сигналу між кінцем поточної паузи та початком наступної паузи.

Схематично алгоритм сегментації звукового сигналу за ознакою зміни диктора подано на рис. 2.

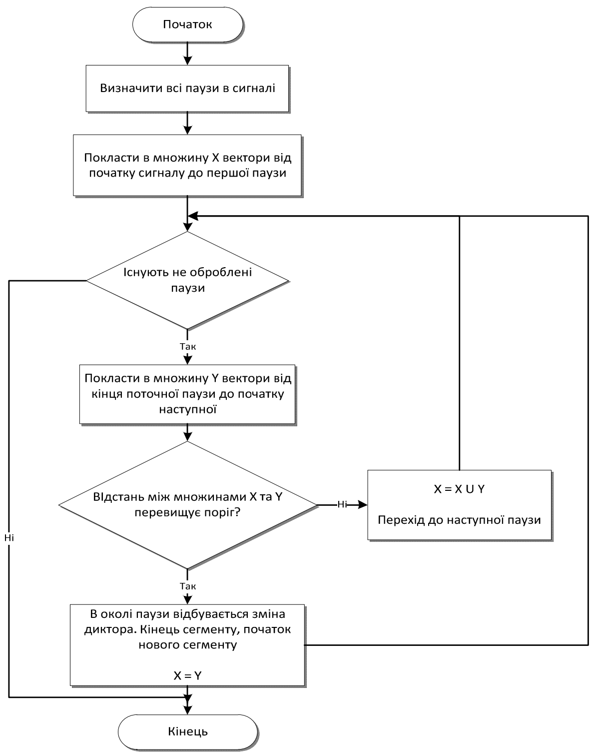


Рисунок 2 – Алгоритм сегментації сигналу за ознакою зміни диктора

Експериментальна перевірка

Описані вище алгоритми визначення пауз і сегментації сигналу за ознакою зміни диктора було реалізовано в системі автоматизованого розподіленого стенографування, де сегментація сигналу для розподілення його між операторами-стенографістами здійснюється за максимальною тривалістю сегмента і ознакою зміни диктора.

Для перевірки алгоритму використовувався запис радіоінтерв'ю, з кількістю змін диктора 27, загальною кількістю пауз, за якими проводилась перевірка зміни диктора, – 298, трьома дикторами, присутніми у записі.

Для перевірки роботи підходу використовувалися метрики, запропоновані в [7], які враховують кількість помилок першого і другого роду:

$$PRC = \frac{\textit{кількість коректно визначених змін}}{\textit{загальна кількість знайдених змін}} \qquad (7)$$

$$RCL = \frac{\textit{кількість коректно визначених змін}}{\textit{загальна кількість змін, присутніх в сигналі}}. \qquad (8)$$

Загальна метрика для порівняння задається як:

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL}. \qquad (9)$$

Така метрика знаходиться в межах між 0 і 1, чим вище її значення, тим краще точність визначення позицій змін диктора.

Результати експерименту наведено в табл. 1. Для порівняння поруч наводяться результати, отримані за допомогою підходу, описаного в [6].

Таблиця 1 – Результати експериментальної перевірки

Підхід	PRC	RCL	F
Запропонований	0,44	0,77	0,56
Медіана відстаней	0,19	0,77	0,30

В обох підходах є достатньо великою кількість помилково визначених змін диктора, що показує відносно невелике значення параметра PRC. Проте в запропонованому підході кількість помилково визначених змін диктора є значно меншою. Причина помилково визначених пауз, як правило, полягає в тому, що підібрати поріг, необхідний для коректної роботи алгоритму, досить складно: занадто низький поріг призводить до великої кількості помилково визначених змін, а занадто високий – до високої кількості пропущених змін диктора.

Деякі зміни диктора в запропонованому підході були також пропущені тому, що при визначенні параметрів нормального розподілу методом максимальної правдоподібності, коваріаційна матриця, яка була отримана в результаті оцінок, не була додатньо визначеною.

Висновки

Запропонований підхід дозволяє достатньо точно визначати зміни дикторів у мовному голосовому сигналі і виконувати сегментацію сигналу за ознакою зміни диктора. За рахунок того, що для при визначенні характеристичних ознак не враховуються сегменти з низькою логарифмічною енергією і пропускаються паузи, вдалося уникнути негативного впливу сторонніх шумів і малоінформативних ділянок сигналу на точність визначення характеристичних ознак диктора.

Подальший розвиток запропонованого підходу має бути направлений на автоматичне визначення порогових значень для міри відмінності між множинами характе-

ристичних векторів, чи позбавлення від необхідності мати порогове значення взагалі, і на вирішення проблеми з коваріаційними матрицями, які отримуються в результаті оцінок за методом максимальної правдоподібності, що не є додатньо визначеними.

Література

1. Автоматизированная система стенографирования / Ю.Г. Кривонос, Ю.В. Крак, А.В. Бармак, А.С. Загваздин // Штучний інтелект. – 2009. – № 3. – С. 228-233.
2. Kotti M. Automatic Speaker Change Detection with the Bayesian Information Criterion using MPEG-7 Features and a Fusion Scheme [Електронний ресурс] / M. Kotti, E. Benetos, C. Kotropoulos // Proc. of ISCAS-2006. – Режим доступу : http://poseidon.csd.auth.gr/papers/PUBLISHED/CONFERENCE/pdf/Kotti06b.pdf
3. Lu L. Speaker change detection and tracking in real-time news broadcasting analysis / L. Lu, H.-J. Zhang // Proceedings of the tenth ACM international conference on Multimedia. – 2002. – P. 602-610.
4. Universal background models for real-time speaker change detection [Електронний ресурс] / T.Y. Wu, L. Lu, K. Chen, H.-J. Zhang // Microsoft Research. – Режим доступу : http://research.microsoft.com/users/llu/publications/mmm03_ubmforspkseg.pdf.
5. Kwon S. Speaker change detection using a new weighted distance measure / S. Kwon, S. Narayanan // Proc. of International conference on spoken language processing. – 2002. – Vol. 4. – P. 2537-2540.
6. Кривонос Ю.Г. Определение позиций изменения диктора в речевом сигнале / Ю.Г. Кривонос, Ю.В. Крак, А.С. Загваздин // Штучний інтелект. – 2010. – № 3. – С. 220-226.
7. Ajmera J. Robust Speaker Change Detection / J. Ajmera, I. McCowan, H. Bourlard // IEEE Signal Processing Letters. – 2004. – № 8, vol. 11. – P. 649-651.
8. Rabiner L. Application of LPC Distance Measure to Voiced-Unvoiced-Silence Detection Problem / L. Rabiner, M. Sambur // IEEE Transaction on Acoustics, Speech and Signal Processing. – 1977. – № 7, vol. 25. – P. 338-343.
9. Tanyer G.S. Voice Activity Detection in Non-stationary Noise / G.S. Tanyer, H. Ozer // IEEE Transactions on Speech and Audio Processing. – 2000. – № 4, vol. 8. – P. 478-482.
10. Загваздін О.С. Автоматичне визначення пауз та зменшення рівня шуму в системі автоматизованого стенографування / О.С. Загваздін // Журнал обчислювальної та прикладної математики. – 2010. – № 2. – С. 35-43.
11. Rabiner L. Applications of non-linear smoothing algorithm to speech processing / L. Rabiner, M. Sambur, C. Schmidt // IEEE Transactions on Acoustics, Speech and Signal Processing. – 1975. – Vol. 23. – P. 552-557.

Literatura

1. Krivonos Ju.G. Shtuchnij intelekt. № 3. 2009. S. 228-233.
2. Kotti M., Benetos E., Kotropoulos C. Automatic Speaker Change Detection with the Bayesian Information Criterion using MPEG-7 Features and a Fusion Scheme. Proc. of ISCAS 2006. http://poseidon.csd.auth.gr/papers/PUBLISHED/CONFERENCE/pdf/Kotti06b.pdf
3. L. Lu. Proceedings of the tenth ACM international conference on Multimediaju. 2002. P. 602-610.
4. T.Y. Wu. Microsoft Research. http://research.microsoft.com/users/llu/publications/mmm03_ubmforspkseg.pdf.
5. S. Kwon, S. Proc. of International conference on spoken language processing. 2002. Vol. 4. P. 2537-2540.
6. Krivonos Ju.G. Shtuchnij intelekt. № 3. 2010. S. 220-226.
7. Ajmera J. IEEE Signal Processing Letters. № 8, vol. 11. 2004. P. 649-651.
8. Rabiner L. IEEE Transaction on Acoustics, Speech and Signal Processing. №7. Vol. 25. 1977. P. 338-343.
9. Tanyer G.S. IEEE Transactions on Speech and Audio Processing. №4. Vol. 8. 2000. P. 478-482.
10. Zagvazdin O.S. Zhurnal obchysljuval'noi ta prikladnoi matematyky. № 2. 2010. S. 35-43.
11. Rabilner L. IEEE Transactions

Yu.G. Kryvonos, Yu.V. Krak, O.S. Zagvazdin, G.M. Yefimov

Speech Signal Segmentation Based on the Speaker Change

An approach to the segmentation of speech signals based on the speaker change, as well as to the detection of the speaker change positions in a speech signal is suggested. Speaker change positions are determined by analyzing the sets of characteristic vectors at the pause within the signal based on the Bayesian information criterion. Improvement in quality of the characteristic vectors is achieved by taking into account only the segments with the log energy above the given threshold. It is also suggested an approach for adaptive automatic pause detection in speech signal.

Стаття надійшла до редакції 22.06.2011.